STA 302f20 Assignment Seven[1]

The following problems are not to be handed in. They are preparation for the Quiz in tutorial and the final exam. Please try them before looking at the answers. Use the formula sheet.

1. Label each of the following statements True (meaning always true) or False (meaning not always true), and show your work or explain. Assume the general linear regression model with normal error terms. As usual, the columns of $\mathbf{X}$ are linearly independent.
   (a) $\widehat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
   (b) $\mathbf{y} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\epsilon}}$.
   (c) $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\epsilon}}$
   (d) $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$
   (e) $\mathbf{X}'\boldsymbol{\epsilon} = \mathbf{0}$
   (f) $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\epsilon}'\boldsymbol{\epsilon}$.
   (g) $\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}} = \mathbf{0}$
   (h) $\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}} = \mathbf{y}'\widehat{\boldsymbol{\epsilon}}$.
   (i) $w = \frac{\boldsymbol{\epsilon}'\boldsymbol{\epsilon}}{\sigma^2}$ has a chi-squared distribution.
   (j) $E(\boldsymbol{\epsilon}'\boldsymbol{\epsilon}) = 0$
   (k) $E(\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}) = 0$

2. For the general linear regression model with normal error terms,
   (a) What is the distribution of the response variable $\mathbf{y}$? Just write down the answer.
   (b) What is the distribution of the vector of estimated regression coefficients $\widehat{\boldsymbol{\beta}}$? Show the calculations.
   (c) What is the distribution of the vector of "predicted" values $\widehat{\mathbf{y}}$? Show the expected value and covariance matrix calculations. Express the covariance matrix in terms of the hat matrix $\mathbf{H}$.
   (d) What is the distribution of the vector of residuals $\widehat{\boldsymbol{\epsilon}}$? Show the calculations. Simplify. Express the covariance matrix in terms of the hat matrix $\mathbf{H}$.

3. For the general linear regression model with normal error terms, show that the $n \times (k+1)$ matrix of covariances $cov(\widehat{\boldsymbol{\epsilon}}, \widehat{\boldsymbol{\beta}}) = \mathbf{O}$. Why does this show that $SSE = \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}$ and $\widehat{\boldsymbol{\beta}}$ are independent?

4. Calculate $cov(\widehat{\boldsymbol{\epsilon}}, \widehat{\mathbf{y}})$; show your work. Why should you have known this answer without doing the calculation, assuming normal error terms? Why does the assumption of normality matter? Is the assumption of normality necessary?

5. What is the distribution of $\mathbf{s}_1 = \mathbf{X}'\boldsymbol{\epsilon}$? Show the calculation of expected value and variance-covariance matrix.

6. What is the distribution of $\mathbf{s}_2 = \mathbf{X}'\widehat{\boldsymbol{\epsilon}}$?

   (a) Answer the question.

   (b) Show the calculation of expected value and variance-covariance matrix.

   (c) Is this a surprise? Answer Yes or No.

   (d) What is the probability that $\mathbf{s}_2 = \mathbf{0}$? The answer is a single number.

7. In an earlier Assignment, you proved that

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}} + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\mathbf{X}'\mathbf{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

   (a) Since you were able to do it once, please do it again for practice. Adding and subtracting the projection is what makes it work.

   (b) Starting with this expression, show that $SSE/\sigma^2 \sim \chi^2(n-k-1)$. Use the formula sheet as necessary.

8. For the general linear regression model with normal errors, tests and confidence intervals for linear combinations of regression coefficients are very useful. Derive the appropriate t distribution and some applications by following these steps. Let $\mathbf{a}$ be a $(k+1) \times 1$ vector of constants.

   (a) What is the distribution of $\mathbf{a}'\widehat{\boldsymbol{\beta}}$? Show a little work. Your answer includes formulas for the parameters of the distribution.

   (b) Now standardize $\mathbf{a}'\widehat{\boldsymbol{\beta}}$ (subtract off the mean and divide by the standard deviation) to obtain a standard normal.

   (c) Divide by the square root of a well-chosen chi-squared random variable, divided by its degrees of freedom, and simplify. Call the result t.

   (d) How do you know numerator and denominator are independent?

   (e) Suppose you wanted to test $H_0 : \mathbf{a}'\boldsymbol{\beta} = c$. Write down a formula for the test statistic.

   (f) For a regression model with four predictor variables, suppose you wanted to test $H_0 : \beta_2 = 0$. Give the vector $\mathbf{a}$.

   (g) For a regression model with four independent variables, suppose you wanted to test $H_0 : \beta_1 = \frac{1}{2}(\beta_2 + \beta_3)$. Give the vector $\mathbf{a}$.

(h) Consider a data set in which there are n first-year students in ECO100. x_1 is High School Calculus mark, x_2 is High School grade point average, x_3 is score on a test of general mathematical knowledge, and y is mark in ECO100. You seek to estimate expected mark for a student with a 91% in High School Calculus, a High School GPA of 83%, and 24 out of 25 on the test. You are estimating $\mathbf{a}'\boldsymbol{\beta}$. Give the vector $\mathbf{a}$.

(i) Letting $t_{\alpha/2}$ denote the point cutting off the top $\alpha/2$ of the t distribution with $n-k-1$ degrees of freedom, derive the $(1-\alpha)\times 100\%$ confidence interval for $\mathbf{a}'\boldsymbol{\beta}$. "Derive" means show the High School algebra.

9. In Question 17 of Assignment 4, you considered a regression model with no predictor variables. If the errors are normal, this is the same as sampling $y_1, \ldots, y_n$ from a normal distribution with expected value $\mu = \beta_0$ and variance σ^2.

(a) What is *MSE* for this problem? Show some work. Feel free to use your results from Assignment 4.

(b) Show that with no predictor variables, the confidence interval for β_0 is just the usual confidence interval for μ. You may use your answer to Question 3 of Assignment 2.

10. For the general linear model with normal errors,

(a) Let $\mathbf{C}$ be a $q \times (k+1)$ matrix of constants with linearly independent rows. What is the distribution of $\mathbf{C}\widehat{\boldsymbol{\beta}}$?

(b) If $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$ is true, what is the distribution of $\frac{1}{\sigma^2}(\mathbf{C}\widehat{\boldsymbol{\beta}}-\mathbf{t})'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\widehat{\boldsymbol{\beta}}-\mathbf{t})$? Please locate support for your answer on the formula sheet. For full marks, don't forget the degrees of freedom.

(c) What is the distribution of SSE/σ^2

i. If H_0 is true?

ii. If H_0 is false?

(d) Form the F ratio for testing $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$.

i. Write down the formula. Simplify.

ii. How do you know numerator and denominator are independent?

iii. What is the distribution of the test statistic if the null hypothesis is true?

11. Suppose you wish to test the null hypothesis that a *single* linear combination of regression coefficients is equal to a scalar constant t_0. That is, you want to test $H_0 : \mathbf{a}'\boldsymbol{\beta} = t_0$. Referring to the formula sheet, verify that $F = t^2$. Show your work.

12. The exact way that you express a linear null hypothesis does not matter. Let $\mathbf{A}$ be a $q \times q$ nonsingular matrix (meaning $\mathbf{A}^{-1}$ exists), so that $\mathbf{C}\boldsymbol{\beta} = \mathbf{t}$ if and only if $\mathbf{AC}\boldsymbol{\beta} = \mathbf{At}$. This is a useful way to express a logically equivalent linear null hypothesis. Show that the general linear test statistic F^* for testing $H_0 : \mathbf{AC}\boldsymbol{\beta} = \mathbf{At}$ is the same as the one for testing $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$.

13. In the full-reduced approach to testing linear null hypotheses, there are several formulas for the test statistic F^*, all connected by a bit of High School algebra.

    (a) Show $\frac{\left(R^2(full) - R^2(reduced)\right)/q}{\left(1 - R^2(full)\right)/(n-k-1)} = \frac{SSR(full) - SSR(reduced)}{q\ MSE(full)}$.

    (b) Show $\frac{SSR(full) - SSR(reduced)}{q\ MSE(full)} = \frac{SSE(reduced) - SSE(full)}{q\ MSE(full)}$

    (c) Show $\frac{SSR(full) - SSR(reduced)}{q\ MSE(full)} = \left(\frac{n-k-1}{q}\right)\left(\frac{p}{1-p}\right)$, where $p = \frac{R^2(full) - R^2(reduced)}{1 - R^2(reduced)}$.

    (d) Show $p = \frac{qF^*}{qF^* + n - k - 1}$.

14. For the general linear regression model with normal error terms, we'd like to show that if the model has an intercept, then $\widehat{\boldsymbol{\epsilon}}$ and $\overline{y}$ are independent. If you can show that $\overline{y}$ is a function of $\widehat{\boldsymbol{\beta}}$, you are done (why?). Here are some ingredients to start you out. For the model with intercept,

    (a) What does $\mathbf{X}'\widehat{\boldsymbol{\epsilon}} = \mathbf{0}$ tell you about $\sum_{i=1}^{n} \widehat{\epsilon}_i$?

    (b) Therefore what do you know about $\sum_{i=1}^{n} y_i$ and $\sum_{i=1}^{n} \widehat{y}_i$?

    (c) Now indicate why $\widehat{\boldsymbol{\epsilon}}$ and $\overline{y}$ are independent.

15. Examine the formulas for $SST = SSE + SSR$ on the formula sheet. How do you know that SSR and SSE are independent if the model has an intercept?

16. Continue assuming that the regression model has an intercept. Many statistical programs automatically provide an *overall* test. The null hypothesis of this test says that none of the predictor variables makes any difference. If you can't reject that, you're in trouble. Supposing $H_0 : \beta_1 = \cdots = \beta_k = 0$ is true,

    (a) What is the distribution of y_i?

    (b) What is the distribution of $\frac{SST}{\sigma^2}$? Just write down the answer. If necessary, check the formula sheet.

    (c) What is the distribution of SSR/σ^2? Use the formula sheet and show your work. Don't forget the degrees of freedom.

    (d) Recall the definition of the F distribution. If $w_1 \sim \chi^2(\nu_1)$ and $w_2 \sim \chi^2(\nu_2)$ are independent, $F = \frac{w_1/\nu_1}{w_2/\nu_2} \sim F(\nu_1, \nu_2)$. Show that $F^* = \frac{SSR/k}{SSE/(n-k-1)}$ has an F distribution under $H_0 : \beta_1 = \cdots = \beta_k = 0$. Refer to earlier results as you use them.

(e) Obtain an F^* test statistic for this same null hypothesis, based on the full-reduced model approach. Does it equal the F^* test statistic of Question 16d?

(f) The null hypothesis $H_0 : \beta_1 = \cdots = \beta_k = 0$ is less and less believable as R^2 becomes larger. Show that the F^* statistic of Question 16d is an increasing function of R^2 for fixed n and k. This means it makes sense to reject H_0 for large values of F.